

Mining Large-Scale Network Data

Jure Leskovec (@jure)

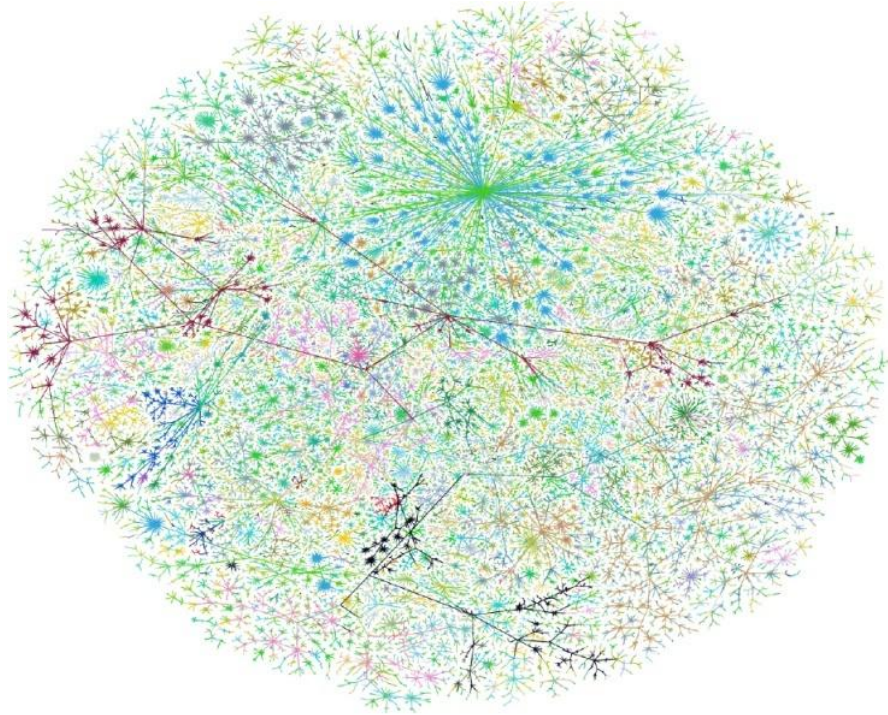
Joint work with L. Backstrom, D. Huttenlocher, J. Kleinberg,
J. McAuley, S. Myers, R. Soric, D. Shahaf, C. Suen



Data Mining & Networks

- **Data mining, Statistics and Machine Learning have rich history and methods for analyzing ...**
 - ... tabular data
 - ... textual data
 - ... time series & streams
 - ... market baskets

Bag of features
- **What about relations and dependencies?**



Networks!

Jure Leskovec, Stanford Computer Forum Annual Meeting, 2013

Networks

are a general language
for reasoning about
real-world systems

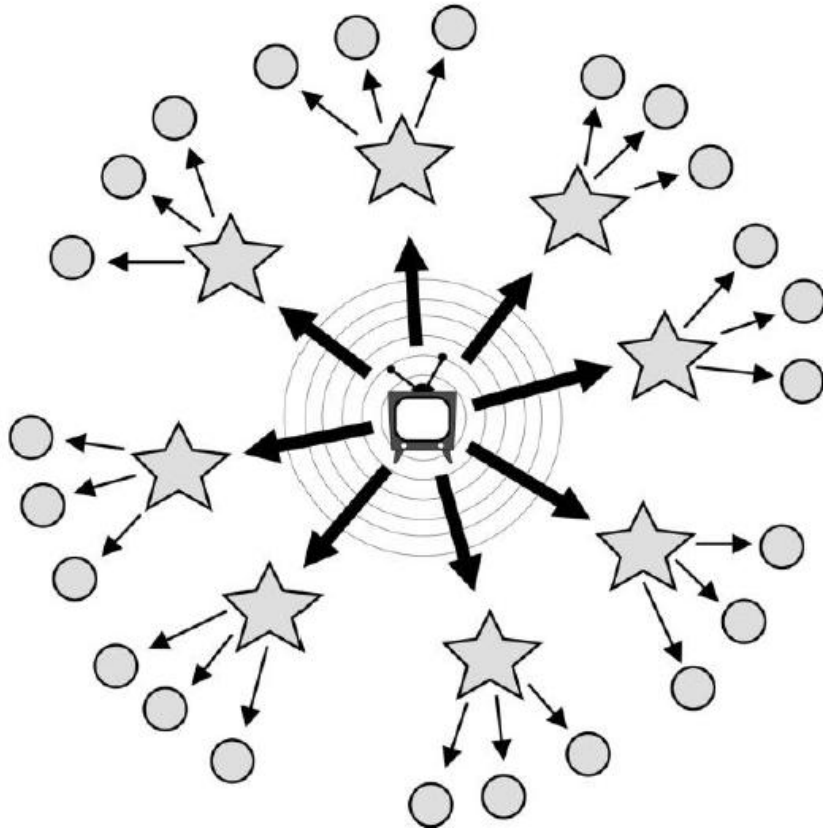


Human interactions



Brain

Jure Leskovec, Stanford Computer Forum Annual Meeting, 2013

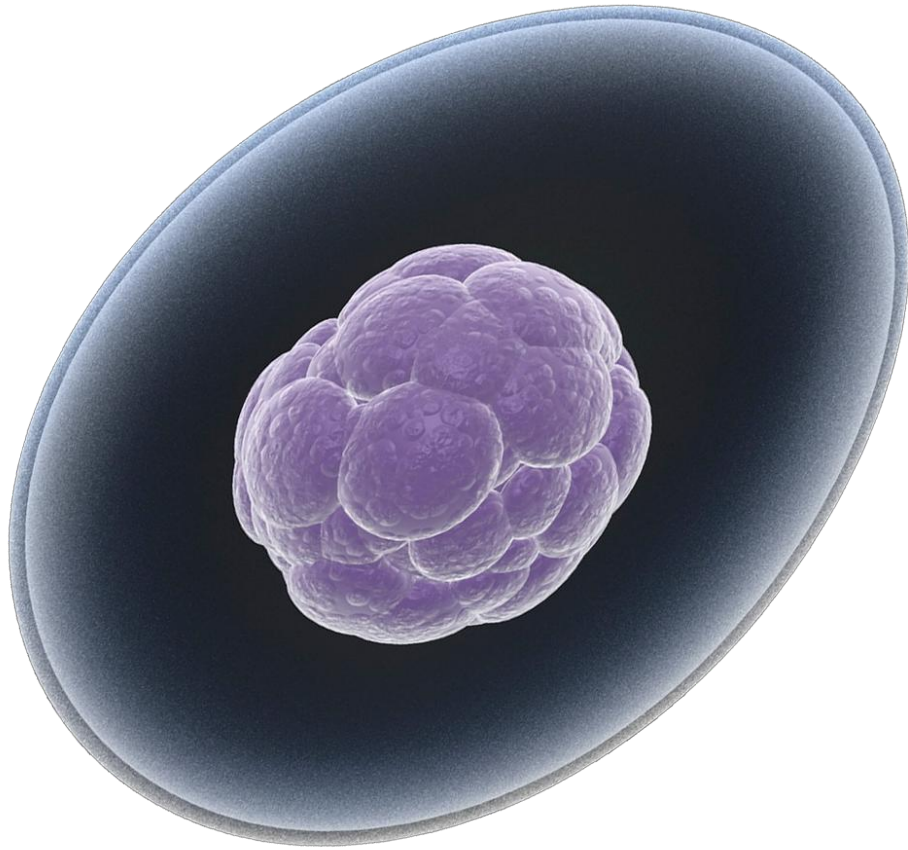


Media & Information



Infrastructure

Jure Leskovec, Stanford Computer Forum Annual Meeting, 2013

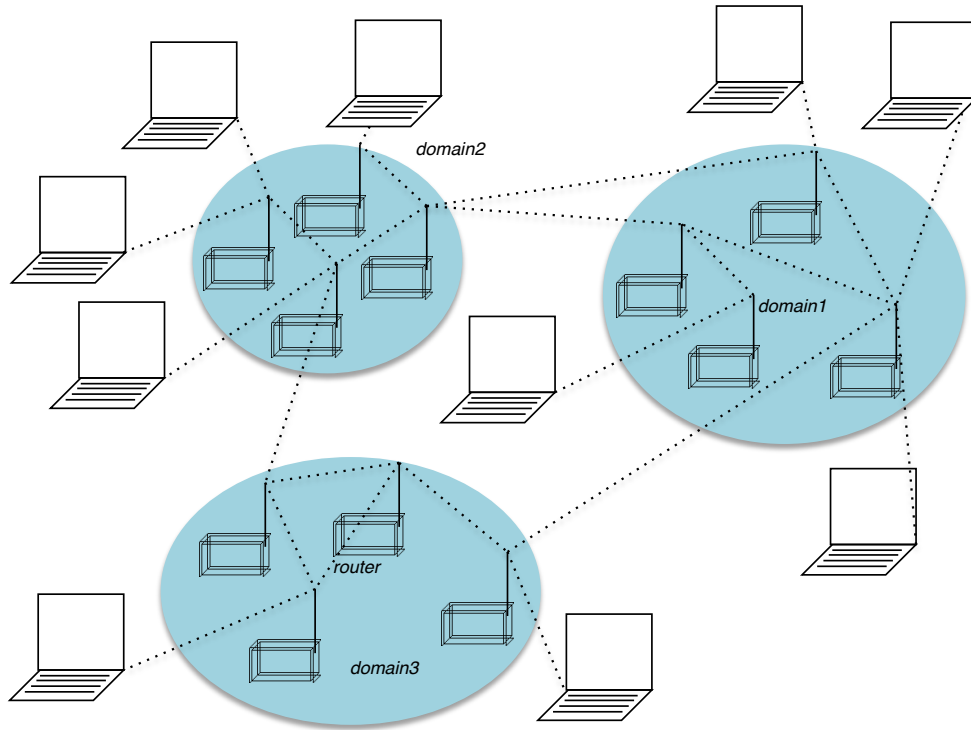


Human cell

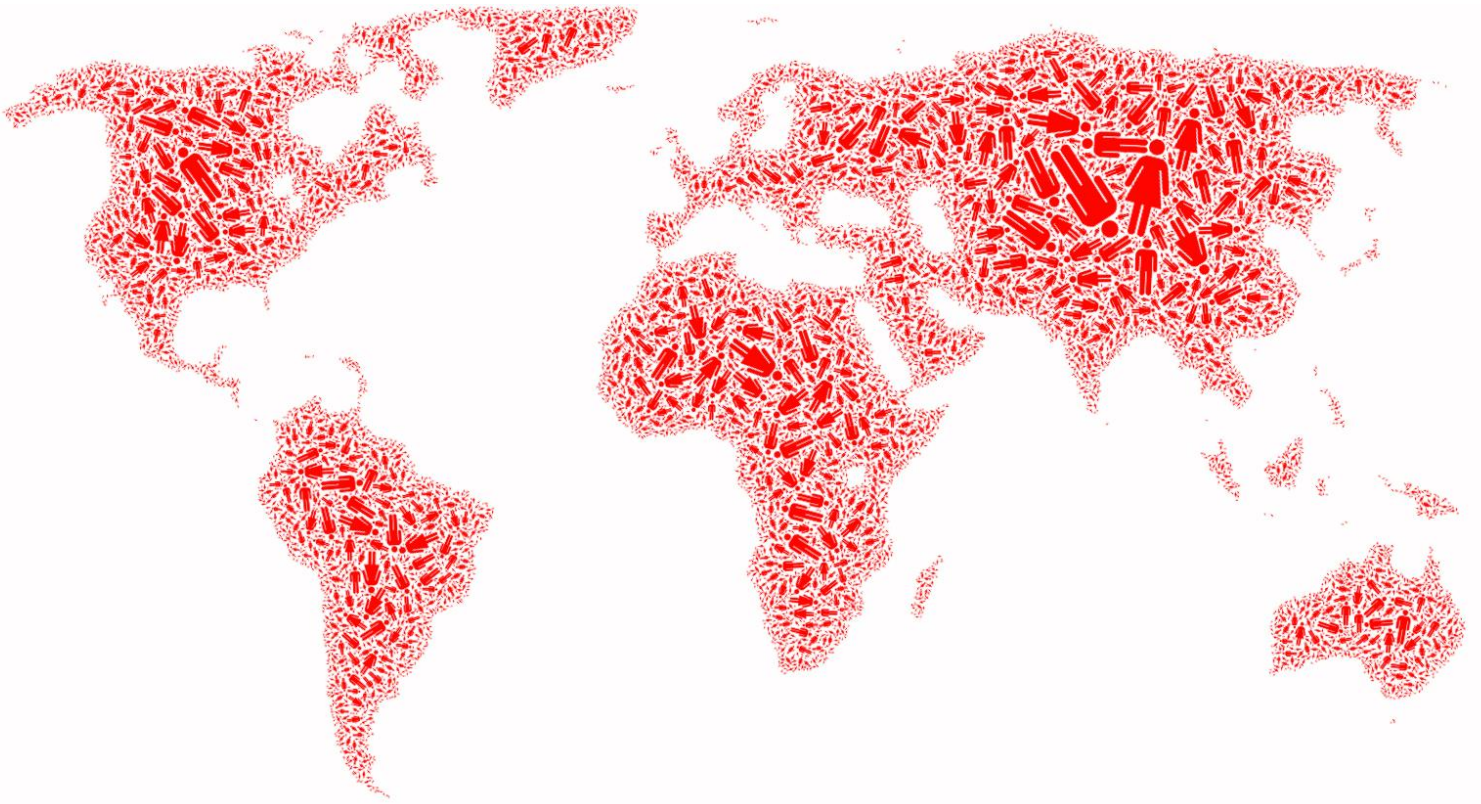


Economy

Jure Leskovec, Stanford Computer Forum Annual Meeting, 2013

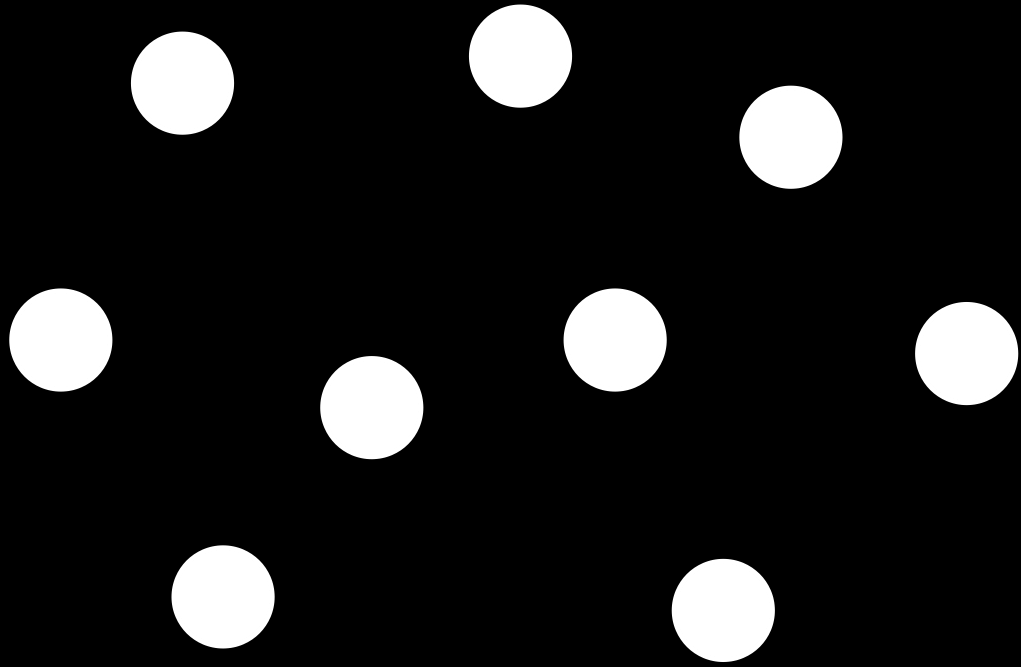


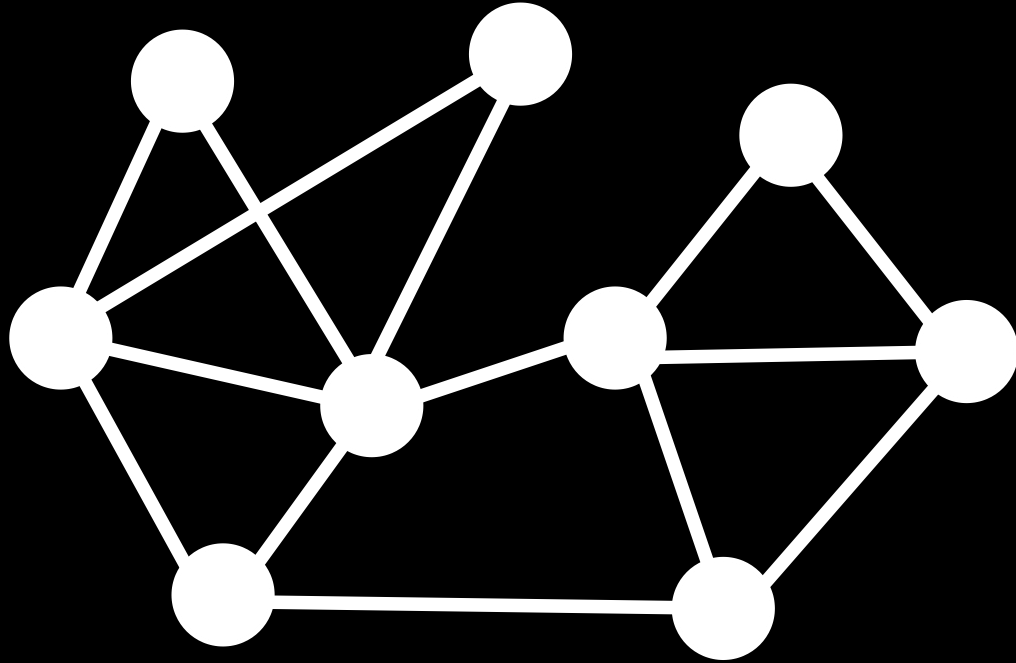
Internet



Society

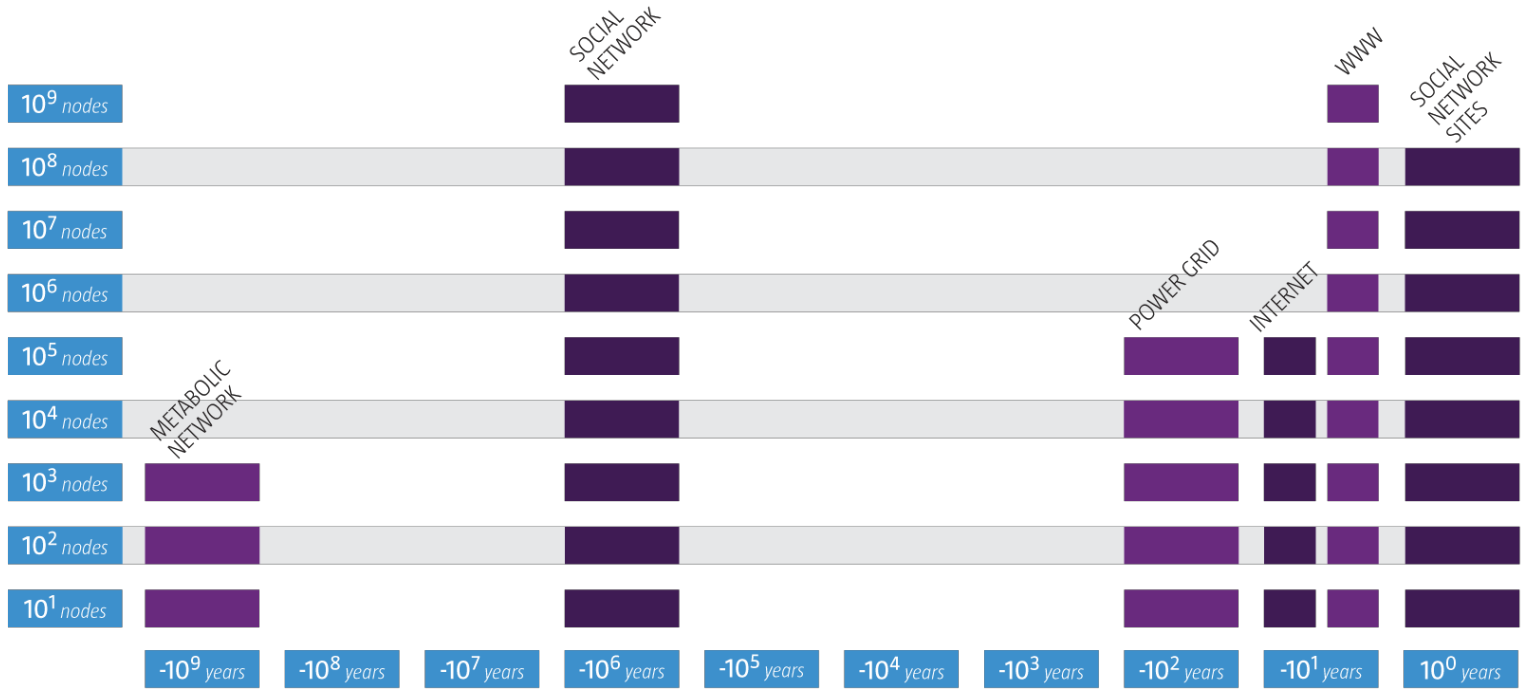
Jure Leskovec, Stanford Computer Forum Annual Meeting, 2013





Network!

Networks, why now?



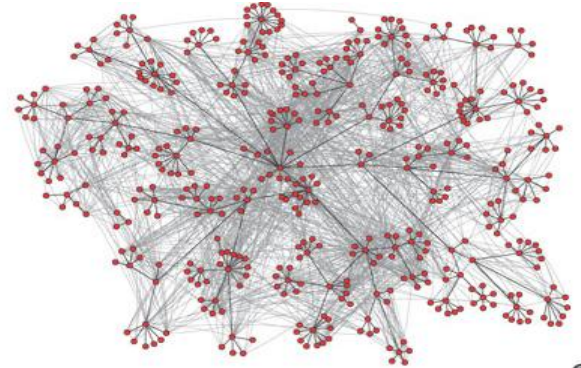
Large-scale network data

Transformation of Computing



Online friendships

[Ugander-Karrer-Backstrom-Marlow, '11]

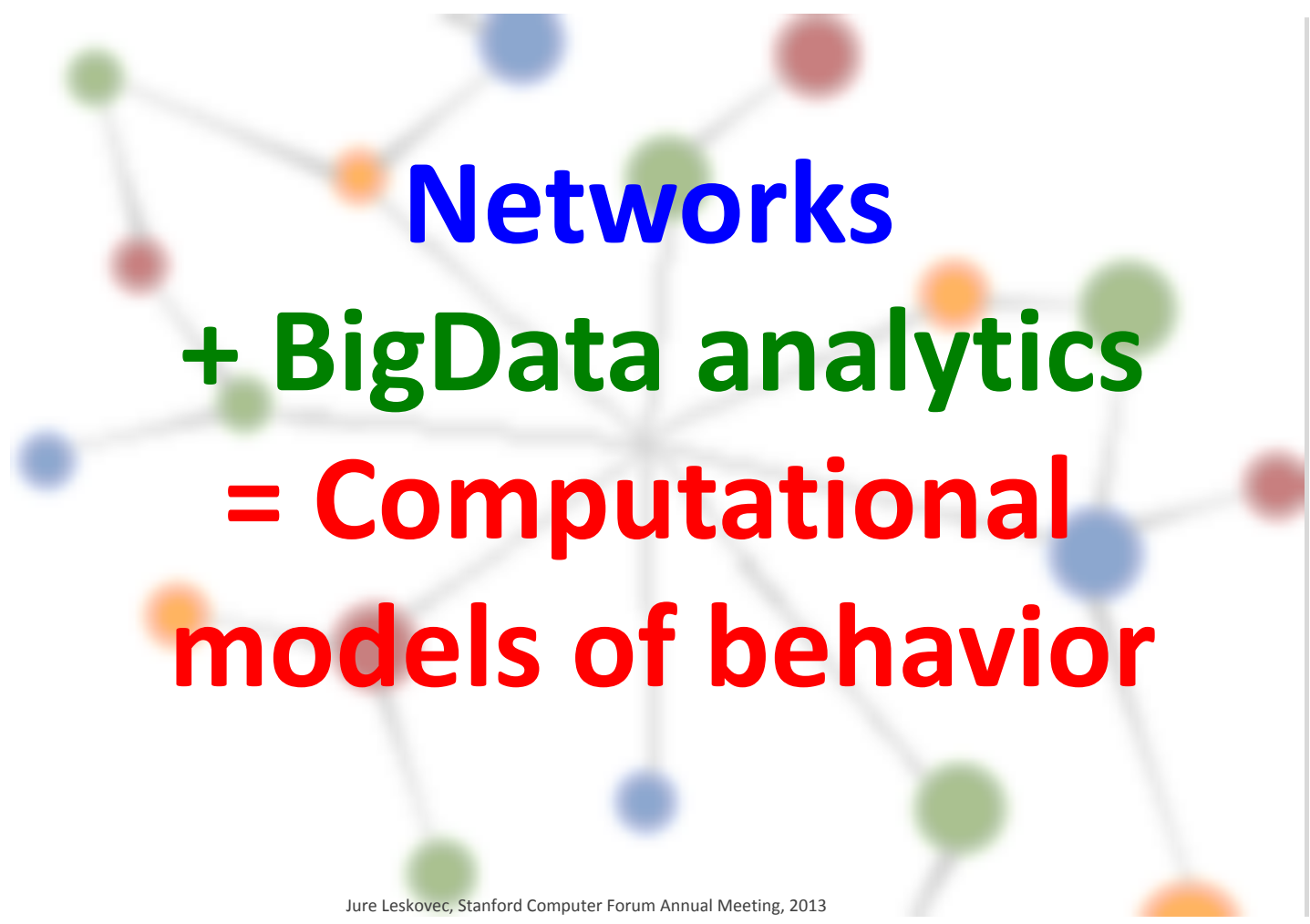


Corporate e-mail communication

[Adamic-Adar, '05]

- **Web is a sensor into humanity!**
- **Profound transformation in:**
 - How knowledge is produced and shared
 - How people interact and communicate
 - **The scope of CS as a discipline**

What do we do?

A background network diagram with nodes of various colors (blue, green, red, orange) connected by thin grey lines. The nodes are scattered across the slide, with a higher density in the center.

Networks
+ BigData analytics
= Computational
models of behavior

Finding Friends

- **Growing body of research captures dynamics of social networks**

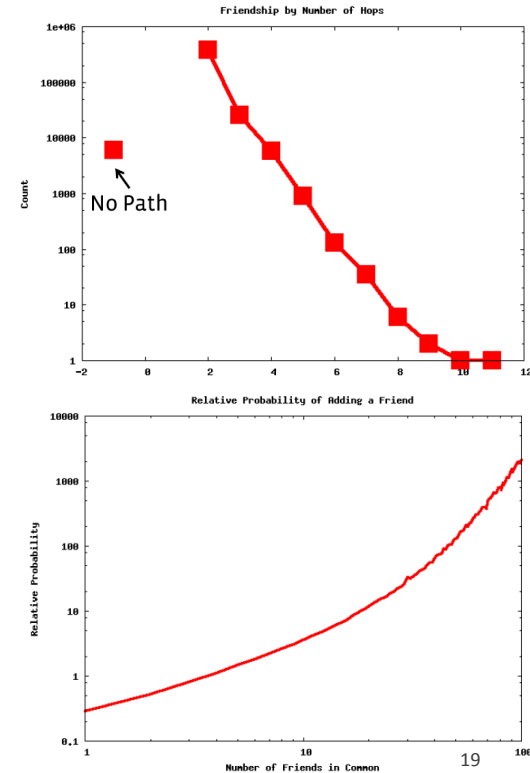
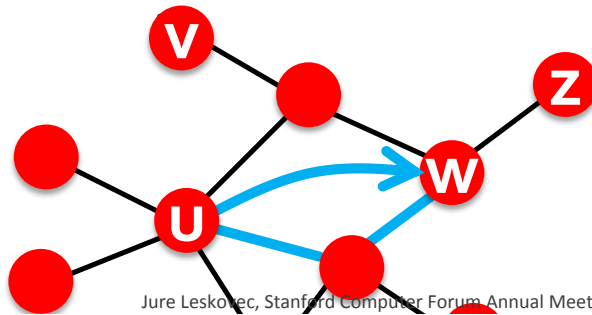
[Latanzi, Sivakumar '08] [Zheleva, Sharara, Getoor '09] [Kumar, Novak, Tomkins '06] [Kossinets, Watts '06] [L., Kleinberg, Faloutsos '05]



- **What links will occur next?** [LibenNowell, Kleinberg '03]
 - **Social network + Many user features:**
 - Location, School, Job, Hobbies, Interests, etc.

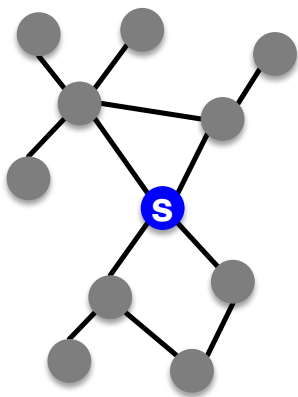
Friend Recommendation

- Learn to recommend friends
- Facebook link creation [Backstrom, L '11]
 - 92% of new friendships on FB are **friend-of-a-friend**
 - **Triadic closure** [Granovetter, '73]
 - More **common friends** helps:
 - **Social capital** [Coleman, '88]

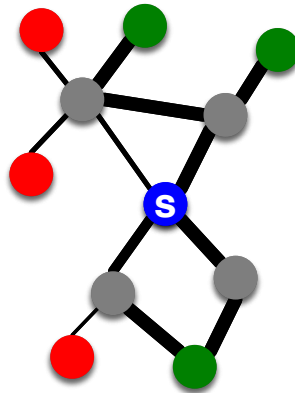
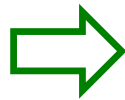


Supervised Link Prediction

- **Goal:** Given a user u , recommend friends
- **Idea:** Learn PageRank scores
 - User features “guide” a random walk



FB Network



Set friendship strengths
(strong edges to point towards **future friends**)



Run Personalized PageRank on a weighted graph



Recommend users with highest score

Link Prediction

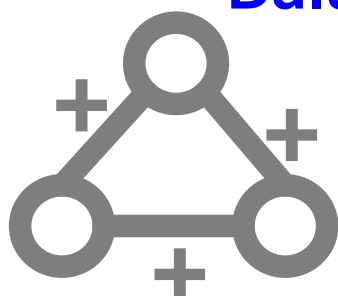
- **Results on Facebook Iceland:**
 - Correctly predicts **8** out of **20 (40%)** new friends
 - 2.3x improvement over previous FB-PYMK



Friend or Foe?

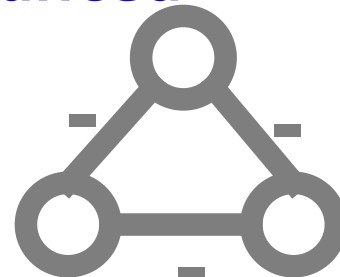
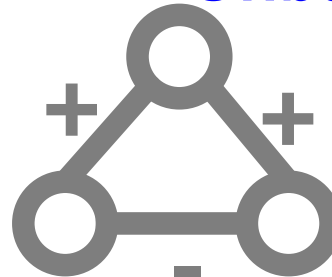
- Not just if you link to someone but also **what do you think of them**
- **Start with the intuition** [Heider '46]
 - The **friend** of my **friend** is my **friend**
 - The **enemy** of **enemy** is my **friend**
 - The **enemy** of **friend** is my **enemy**
 - The **friend** of my **enemy** is my **enemy**

Balanced

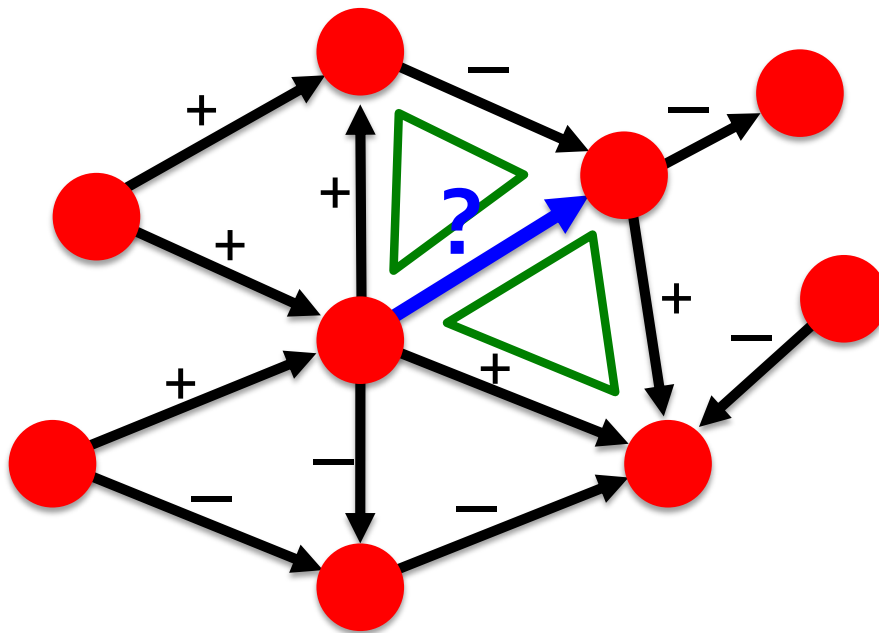


Jure Leskovec, Stanford Computer Forum Annual Meeting, 2013

Unbalanced



Friend or Foe?



> 90% accuracy

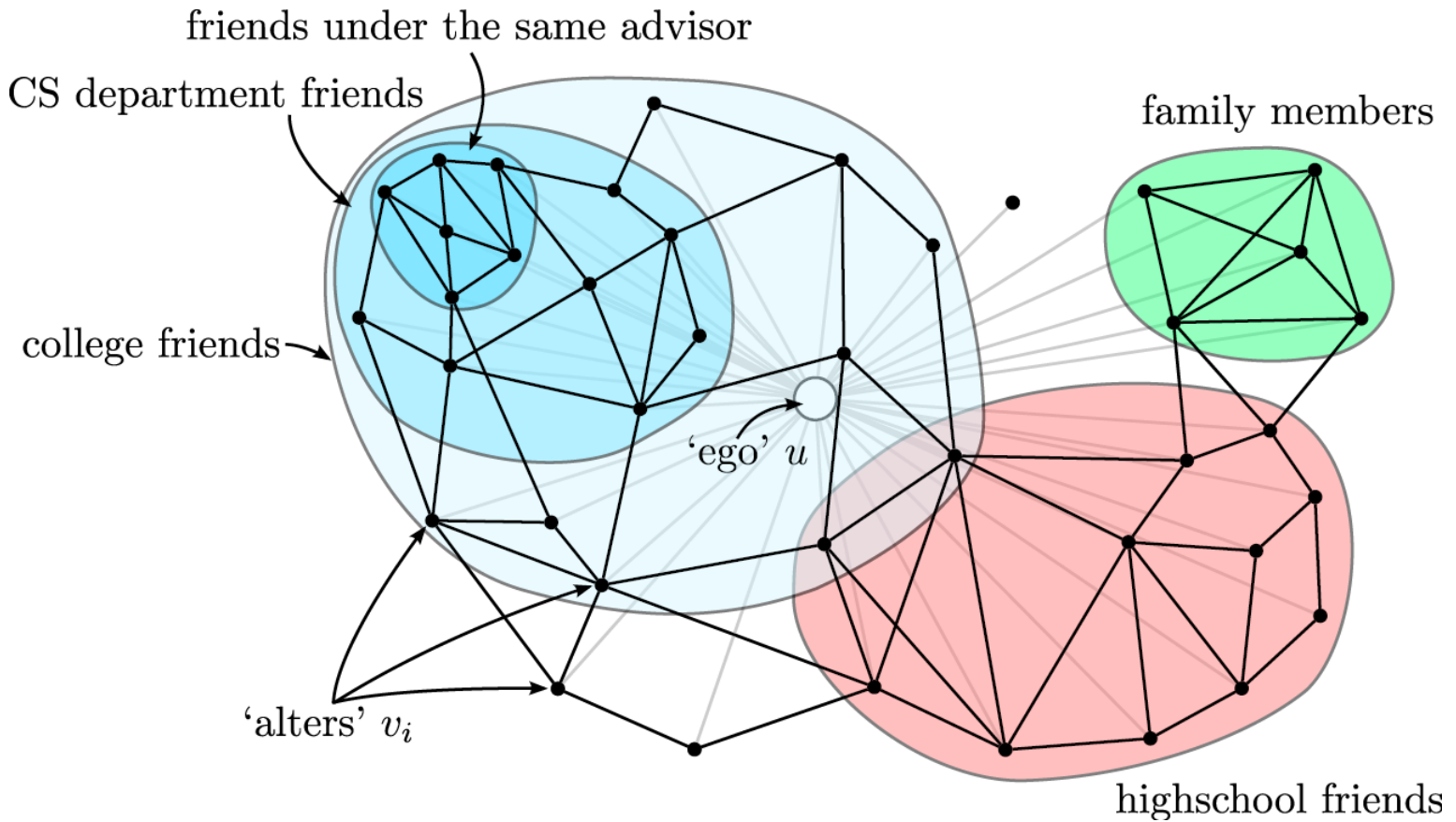
■ Why organize friends?

- Filter and organize content
- Control privacy and access



■ All social networks have this feature:

- Facebook (groups), Twitter (lists), G+ (circles)
- **But circles have to be created manually!**



Discover circles and why they exist

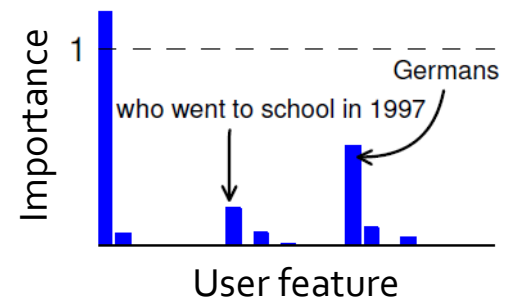
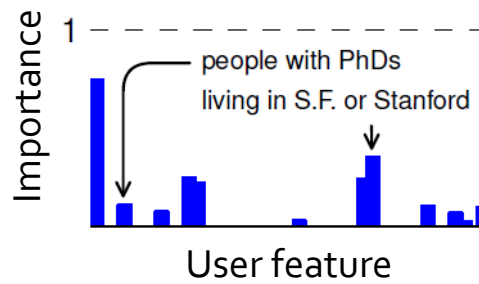
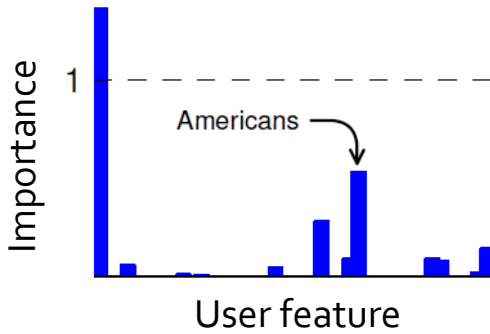
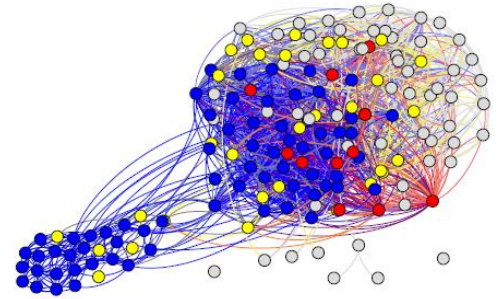
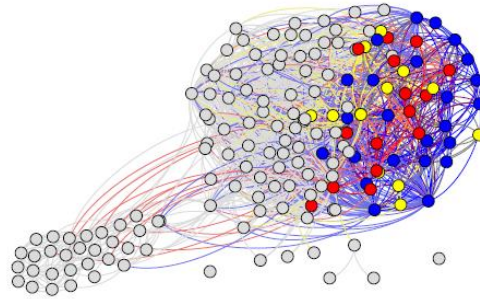
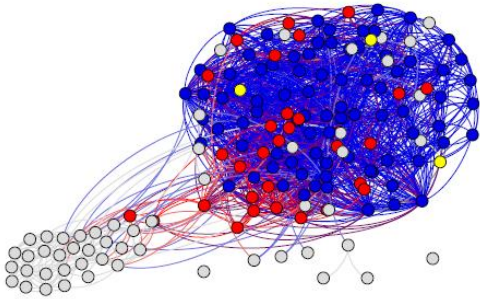
Model of Social Circles

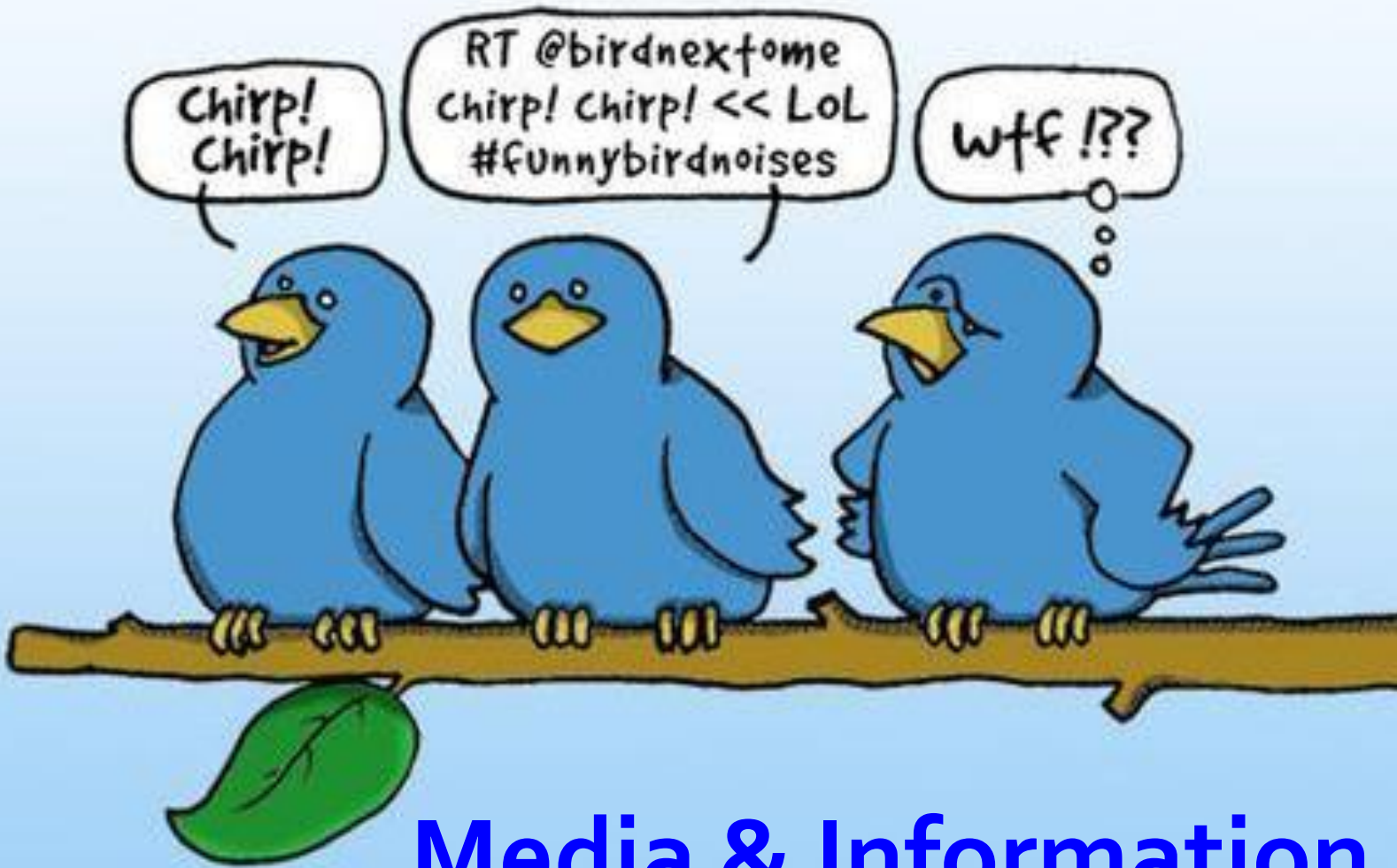
- Suppose we know all the circles
- For a set of circles \mathcal{C} model edge prob.:
 - $(i, j) \propto \exp(-\sum_{c \in \mathcal{C}} \phi_c(i, j))$
 - (i, j) ...is edge feature vector describing (i, j)
 - ϕ_c ...circle parameters that we aim to estimate
- **Example:**

$$(i, j) = \begin{bmatrix} 1 \\ 1 \\ 0 \\ 1 \\ 1 \\ 0 \\ 0 \end{bmatrix} \begin{array}{l} \text{work : position : Cryptanalyst} \\ \text{work : location : GC\&CS} \\ \text{work : location : Royal Navy} \\ \text{education : name : Cambridge} \\ \text{education : type : College} \\ \text{education : name : Princeton} \\ \text{education : type : Graduate School} \end{array} = \begin{bmatrix} \cdot \\ \cdot \\ \cdot \\ \cdot \\ \cdot \\ \cdot \\ \cdot \end{bmatrix}$$

Jure Leskovec, Stanford Computer Forum Annual Meeting, 2013

- How well do we recover human circles?
- Social circles of a particular person:

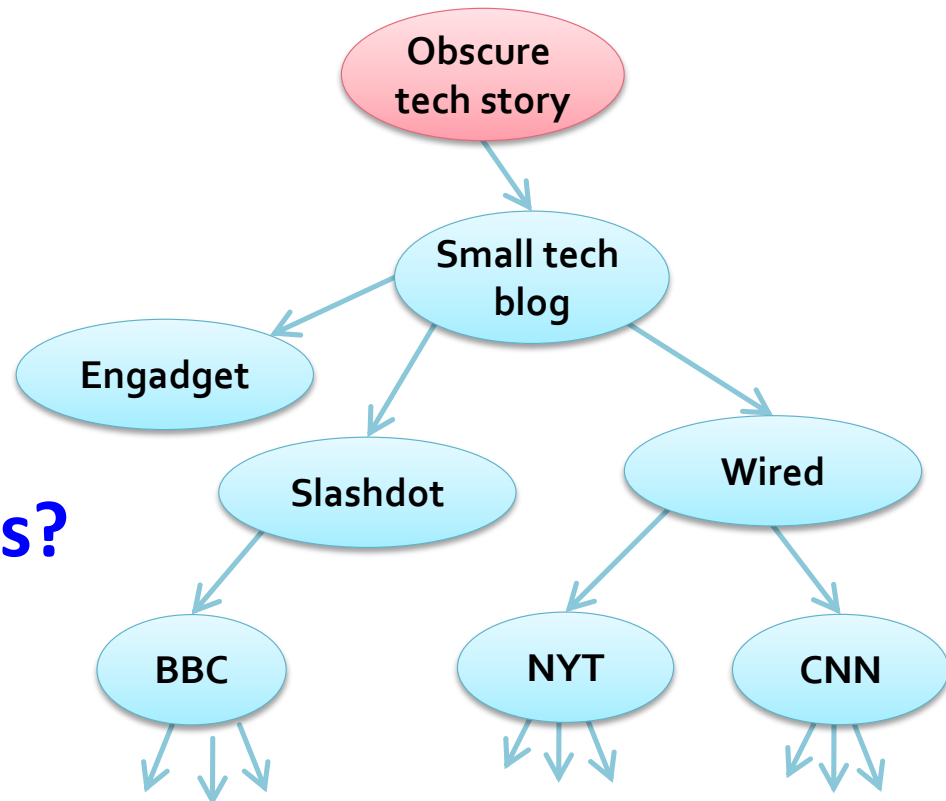




Media & Information

Information in Networks

- How does information interact with our personal social networks?



- Information flows from a node to a node like an epidemic

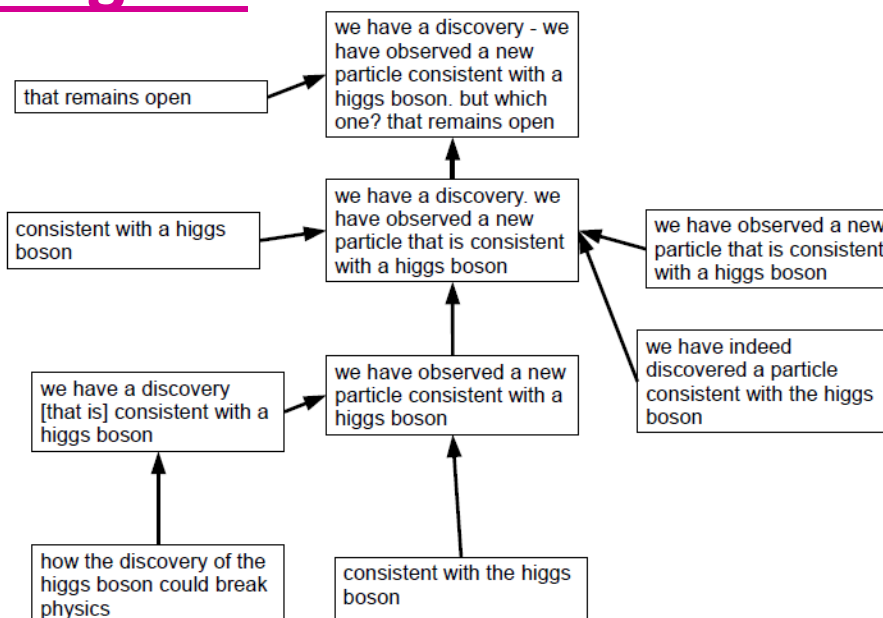
Online Media



- Since August 2008 we have been collecting 30M articles/day: 6B articles, 20TB of data
- Challenge:
How to track information as it spreads?

Meme-tracking

- **Goal:** Trace textual phrases that spread through many news articles
- **Challenge 1: Phrases mutate!**

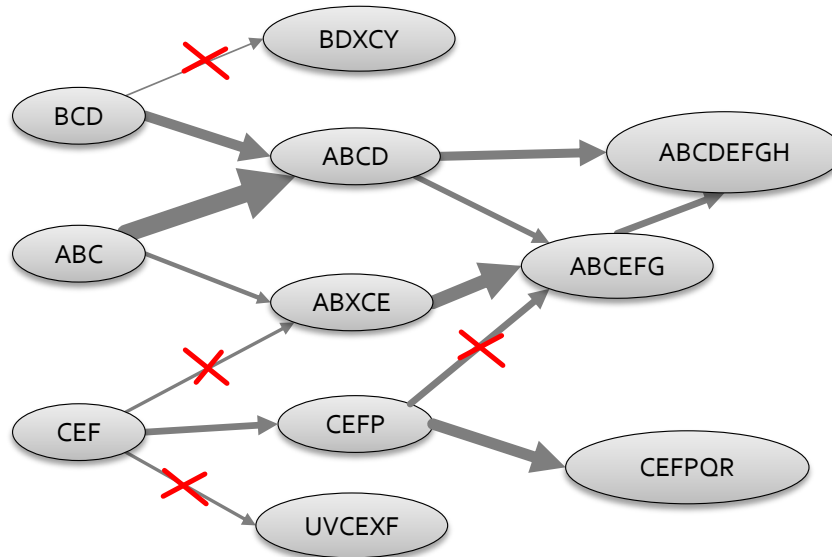


Mutations of a phrase about the Higgs boson particle.

Jure Leskovec, Stanford Computer Forum Annual Meeting, 2013

Finding Mutational Variants

- **Goal:** Find mutational variants of phrases



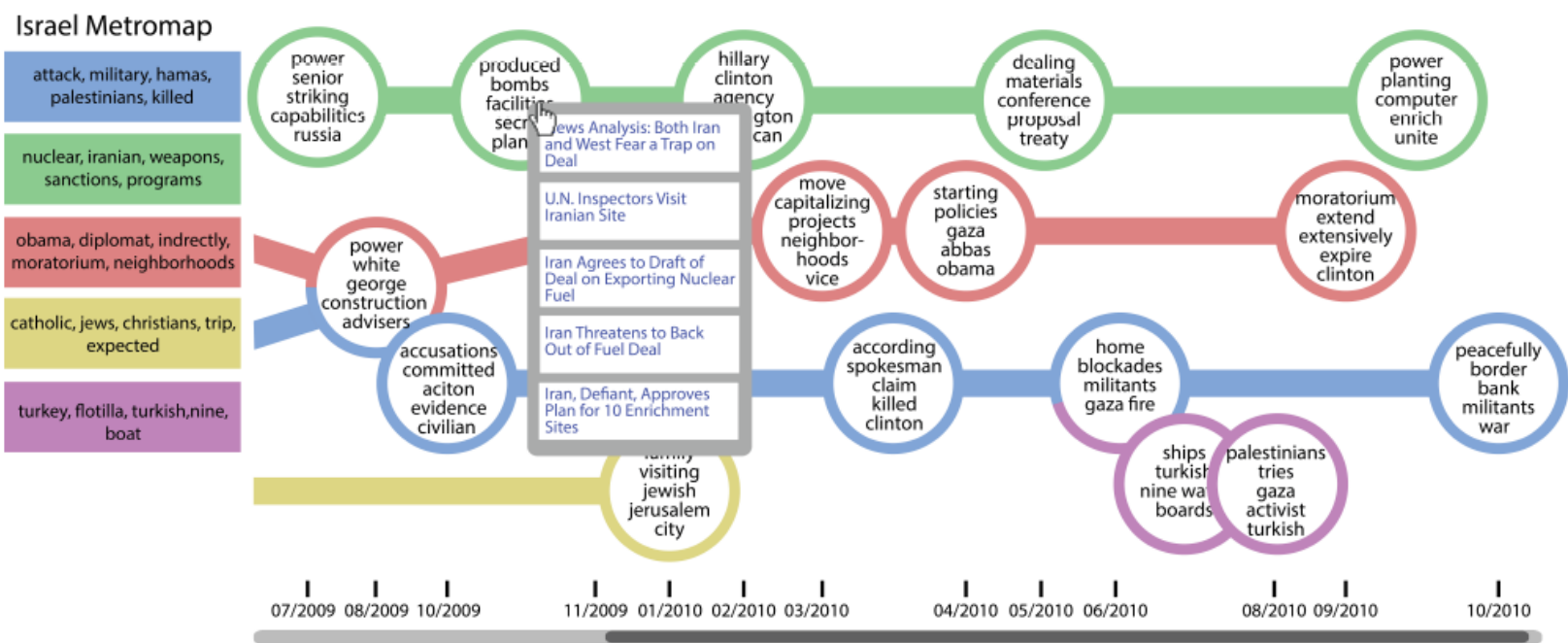
Nodes are phrases
Edges are inclusions
Edges have weights

- **Challenge 2: 20TB of data!**
- **Solution: Incremental partitioning**

New Interaction Techniques

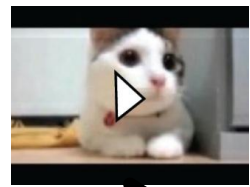


Visualization of 1 month of data from October 2012

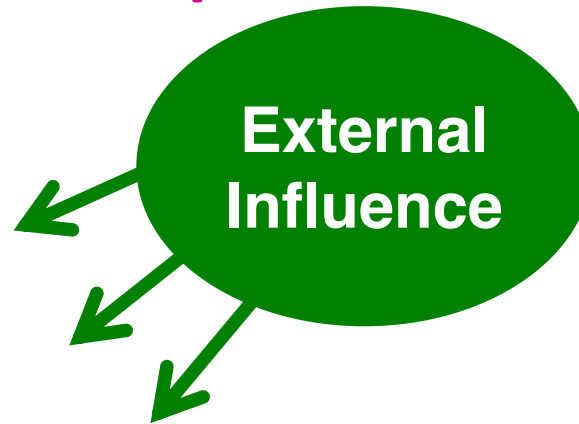


MetroMap of "Israel"

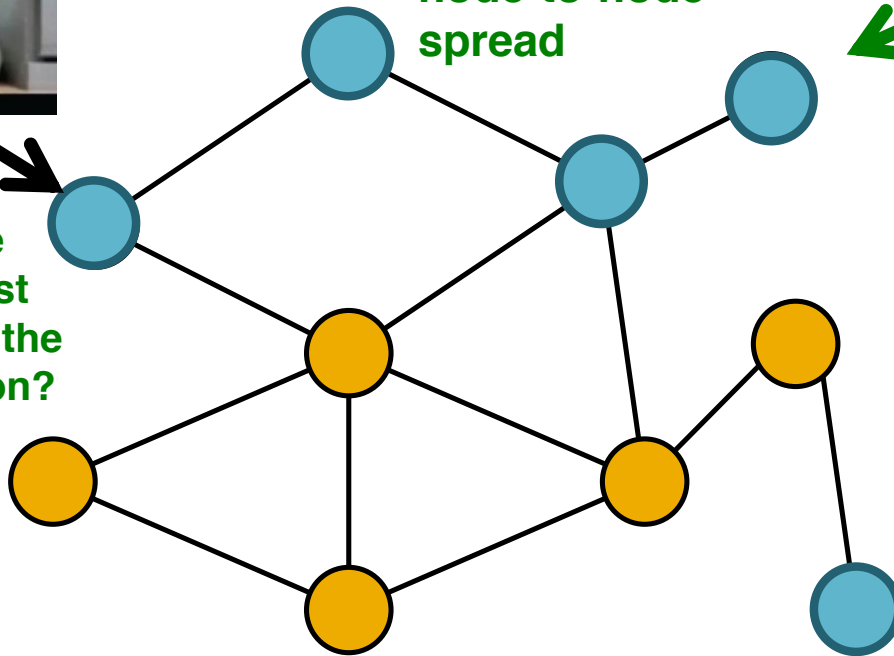
- Observe times when nodes adopt the information



Potential
node-to-node
spread



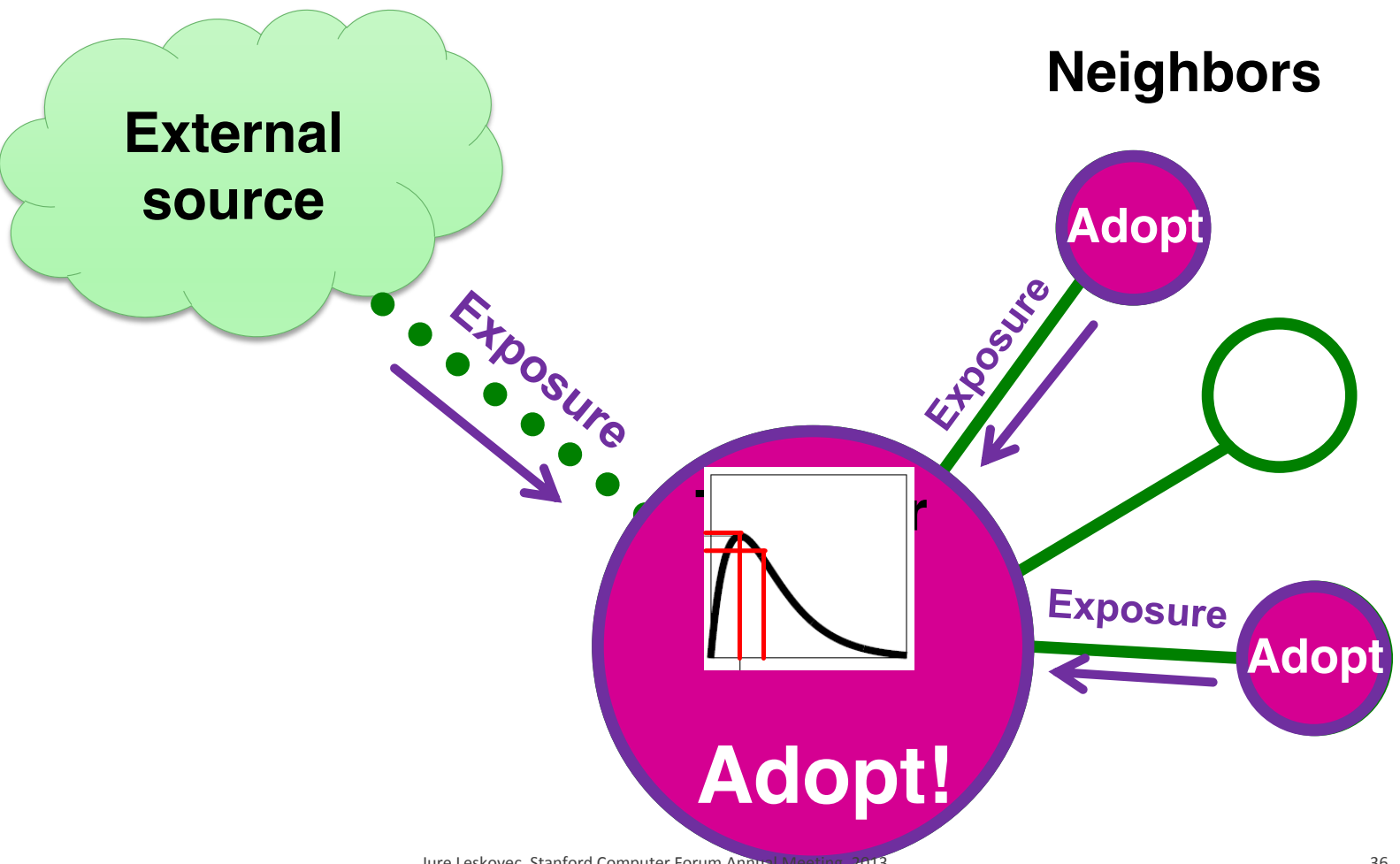
But where
did the first
node find the
information?



How did the
information
“jump”?

Information Adoption

Poster

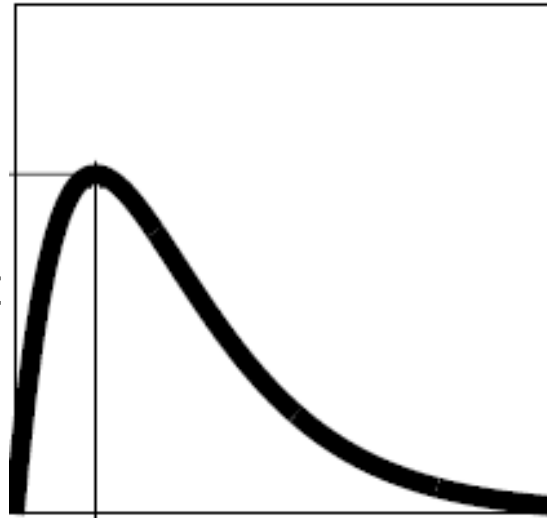


Results: Twitter

Poster

Most infectious:
Business, Entertain.

Least infectious:
Travel, Art



**Need few
exposures:**
Travel, Tech

**Need many
exposures:**
Art, Science

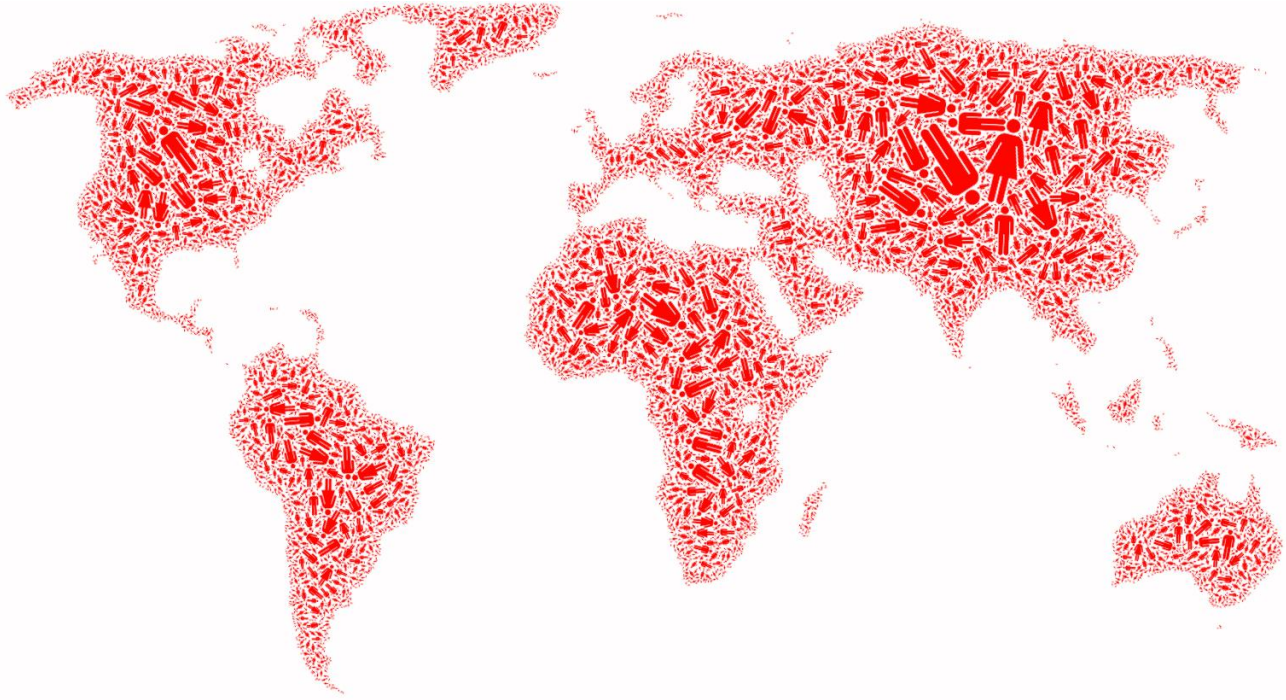
Socially driven:
Tech, Health, Entertain

Externally driven:
Sports, Politics

More details: Myers, Zhu, L. : Information diffusion and external influence in networks, *KDD* 2012.

Jure Leskovec, Stanford Computer Forum Annual Meeting, 2013

What's beyond?



**Networks are a natural language
for reasoning about problems spanning
society, technology and information**

Jure Leskovec, Stanford Computer Forum Annual Meeting, 2013

38

Conclusion & Reflections

- **Only recently has large scale network data become available**
- Opportunity for large scale analyses
- **Benefits of working with massive data**
 - Observe “invisible” patterns

Towards the Model of You

- **Social networks — implicit for millennia — are being recorded in our information systems**
- **Software has a complete trace of your activities — and increasingly knows more about your behavior than you do**
- **Models based on algorithmic ideas will be crucial in understanding these developments**

A screenshot from the game Eve Online showing a vast space station or city built on a planet's surface. The foreground is filled with rows of industrial buildings and structures. In the sky, numerous spaceships of various sizes are flying, some with bright lights. A large, bright sun is visible on the horizon, creating a lens flare effect. The planet's curvature is visible in the distance.

THANKS!

Data + Code:

<http://snap.stanford.edu>

@jure

References

- [Supervised Random Walks: Predicting and Recommending Links in Social Networks](#) by L. Backstrom, J. Leskovec. ACM International Conference on Web Search and Data Mining (WSDM), 2011.
- [Predicting Positive and Negative Links in Online Social Networks](#) by J. Leskovec, D. Huttenlocher, J. Kleinberg. ACM WWW International conference on World Wide Web (WWW), 2010.
- [Learning to Discover Social Circles in Ego Networks](#) by J. McAuley, J. Leskovec. Neural Information Processing Systems (NIPS), 2012.
- [Defining and Evaluating Network Communities based on Ground-truth](#) by J. Yang, J. Leskovec. IEEE International Conference On Data Mining (ICDM), 2012.
- [The Life and Death of Online Groups: Predicting Group Growth and Longevity](#) by S. Kairam, D. Wang, J. Leskovec. ACM International Conference on Web Search and Data Mining (WSDM), 2012.

References

- [Meme-tracking and the Dynamics of the News Cycle](#) by J. Leskovec, L. Backstrom, J. Kleinberg. ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD), 2009.
- [Inferring Networks of Diffusion and Influence](#) by M. Gomez-Rodriguez, J. Leskovec, A. Krause. ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD), 2010.
- [On the Convexity of Latent Social Network Inference](#) by S. A. Myers, J. Leskovec. Neural Information Processing Systems (NIPS), 2010.
- [Structure and Dynamics of Information Pathways in Online Media](#) by M. Gomez-Rodriguez, J. Leskovec, B. Schoelkopf. ACM International Conference on Web Search and Data Mining (WSDM), 2013.
- [Modeling Information Diffusion in Implicit Networks](#) by J. Yang, J. Leskovec. IEEE International Conference On Data Mining (ICDM), 2010.
- [Information Diffusion and External Influence in Networks](#) by S. Myers, C. Zhu, J. Leskovec. ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD), 2012.
- [Clash of the Contagions: Cooperation and Competition in Information Diffusion](#) by S. Myers, J. Leskovec. IEEE International Conference On Data Mining (ICDM), 2012.